ESTHER DYSON'S MONTHLY REPORT

Release 1.0®

VOLUME 19, NO. 11 | 14 DECEMBER 2001 | www.edventure.com

Realtime Computing:
Beyond the Refresh Button

BY KEVIN WERBACH

INSIDE

REALTIME COMPUTING 1
Let's Get Real 2
The need for speed
Realtimers 3
Bang Networks
Kenamea
KnowNow
Realtime is on Our Side 8
The competitive neighborhood
The right time for realtime

NEW ADVENTURES IN TEXT PROCESSING 11
Sense vs. search
Resumé masters
The Sense-Makers 14
WhizBang Labs
Engenium
Burning Glass
Box: Classification and Taxonomies
Mohomine
Box: Caesius
The New Information Toolbox 20
Resources & Contact Information 21
Calendar of High-Tech Events 22

Sign up now for the 2002 PC Forum (our 25th anniversary)! See pp. 23-24 for details and a registration form.

The refresh button's days may be numbered. Realtime enterprise computing promises that business applications and processes running across the Internet will deliver the performance of software on local desktops or on tightly controlled private networks. Using these mechanisms, organizations will become more responsive and better able to link their activities into dynamic webs of business relationships.

As companies deploy Web-services architectures to link applications and systems (SEE **RELEASE 1.0**, OCTOBER AND NOVEMBER 2001), they and their users will want the same experience they are used to with traditional tightly coupled systems. HTML interfaces are just the front end. A true interactive application needs an asynchronous event-driven messaging environment behind it, which means replacing or augmenting the request/response transport paradigm of the Web's hypertext transfer protocol (HTTP).

(We're not talking here about realtime operating systems in embedded devices or other types of realtime software. Realtime enterprise computing refers to the ends, not the means. The data flows and application interfaces appear in the user's "realtime," even though underneath is the non-deterministic Internet.)

The Web uses a page metaphor, which works great for documents but breaks down quickly for applications. The differences in experience between a Website and a "real" application are obvious to anyone who has used both. Applications respond immediately to your actions, while Web pages take time to reload whenever something changes. The Web is not so different from a batch-processing sys-

{ *continued on page 2* }

tem: Users enter data on a page or click on a link, and then initiate a request to a server, which sends back a response.

Existing mechanisms to add interactivity to Web browsers have their own problems. Java applets and dynamic HTML scripting have performance issues and can cause compatibility issues. Because they operate on the client, they don't address the fundamental limitations in the way browsers communicate with servers. Whenever new information comes from the server or another client feeding into the application, the user must request a refresh.

Simply put, HTTP isn't designed for reliable interactive connections. Every update, even if it only changes one line of text, requires a refresh of the whole page. And there's no guarantee a response will come back at all. The browser that initiates the request has no control over the network connecting it to the server.

Let's Get Realtime

The critical elements of realtime computing are interactivity, reliability and security. Interactivity here means instant responsiveness and the ability to act on information immediately. Think of the difference between a live auction, where you see bids as they happen, and one on eBay. Or compare a logistics application that shows trucks moving around a city compared to a series of snapshot maps updated every few minutes. Reliability and security are familiar requirements for enterprise apps; the difficulty lies in meeting those requirements at the speed and scale necessary for true interactivity.

The Internet gave up these three characteristics in exchange for its tremendous simplicity, compatibility and extensibility. The Net uses "best-efforts" transport because that way no provider need be responsible for the burden of ensuring a packet gets from point A to point B. But when you're operating a trading floor or a supply chain, best-efforts delivery isn't good enough. The challenge is to marry realtime qualities with the best aspects of the Internet.

Release 1.0® (ISSN 1047-935X) is published monthly except for a combined July/August issue by EDventure Holdings Inc., 104 Fifth Avenue, New York, NY 10011-6987; 1 (212) 924-8800; fax, 1 (212) 924-0240; www.edventure.com. It covers software, the Internet, e-commerce, convergence, online services, Web applications, technology policy, groupware, data networking, streaming media, enterprise applications, wireless communications, intellectual property and other unpredictable topics.

EDITOR-IN-CHIEF: Esther Dyson (edyson@edventure.com)

PUBLISHER: Daphne Kis (daphne@edventure.com)

EDITOR: Kevin Werbach (kevin@edventure.com)

SYSTEMS DIRECTOR: Scott Giering (scott@edventure.com)

CIRCULATION MANAGER: Natasha Felshman (natasha@edventure.com)

OFFICE MANAGER: Bill Demeritt (bill@edventure.com)

EXECUTIVE ASSISTANT: Irene Lawrence (irene@edventure.com)

CONSULTING EDITOR: Bill Kutik (bill@kutik.com)

Copyright © 2001, EDventure Holdings Inc. All rights reserved. No material in this publication may be reproduced without written permission; however, we gladly arrange for reprints, bulk orders or site licenses. Subscriptions cost $795 per year in the US, Canada and Mexico; $850 overseas.

The need for speed

Realtime computing generates several benefits. In an increasingly digital, increasingly interconnected world, competitive advantage becomes difficult to maintain. Speed and insight become the most valuable tools an organization can possess. Speed requires ever-greater levels of automation, because whenever a task can be reduced to a formula, chances are computers can do it faster and more reliably than people.

The rise of back-office enterprise resource planning (ERP) and front-office customer relationship management (CRM) as the over-arching enterprise software categories reflects this desire for automation. But it doesn't stop at the walls of the enterprise. Supply chain and business-to-business transactional systems, though less developed, are bringing the same kind of automated connectivity to data flows between companies.

Sometimes, speed *is* insight. Stare at a flower and you won't notice it growing. But watch a time-lapse movie of the same plant and you can watch it go from seed to mature flower. More to the point, understanding that your inventory is excessive after the fact doesn't get you very far. You want to see the imbalance as it's happening, and have the tools to correct it. . . or even predict it!

In other words, though realtime computing involves machines, it becomes valuable only when it can get information to *people* who can make use of it through good monitoring systems. The Internet and the Web browser offer a standard interface through which users can take advantage of realtime applications. What's missing is the transport to get the data there immediately and reliably at Internet scale.

Three venture-backed Bay Area startups hope to deliver the next-generation infrastructure to solve this problem: Kenamea, Bang Networks and KnowNow. We discuss each, and several companies pursuing related areas, below.

Realtimers

Bang Networks, Kenamea and KnowNow all launched in 2001, and though they initially focused on different markets, their positioning has converged significantly. Despite the tough market for startups, all three companies raised second financing rounds in 2001, and all are now picking up customers and revenue.

Each of the three is a second-generation Internet-based enterprise software company. Bang draws several key executives from Inktomi. KnowNow's senior management comes from WebMethods, Netscape and Bowstreet. Kenamea's cofounder was chief architect of WebLogic, where he helped develop the dominant e-business application server platform.

Those ties influence each company's strategy and technology, though all share a vision of Internet-based applications that deliver desktop-level performance. Says Bang Networks co-founder and cto Tim Tuttle: "What everyone wants to move to now is applications that have all the functionality of enterprise applications, but they can do it on the Web."

Today, on their own managed networks, enterprises use messaging buses such as Tibco and IBM's MQ Series for their realtime applications. The three startups hope to offer something similar for applications that use the Internet. That means both that they must handle large numbers of endpoints and that they must deal with the Net's unpredictable connectivity and still provide reliable performance. Kenamea ceo John Blair draws an analogy to the difference between IBM's CICS and BEA's WebLogic: Both are forms of online transaction processing, but WebLogic was designed for the needs of Internet applications.

Internet-scale messaging involves thousands of message queues which are short-lived and must operate over a heterogeneous, uncontrolled network. That's a far cry from the traditional enterprise messaging environment, which usually has a small number of queues delivering large amounts of data over tightly controlled networks. Because of these different environments, the startups don't compete directly with Tibco and IBM today, though in the future they may.

Bang Networks: more bang for your financial services

Bang Networks has built its own network infrastructure for reliable, scalable real-time Internet messaging. The original core intellectual property, which co-founder Tuttle developed over a two-year period with a team at MIT's Artificial Intelligence lab, is what Bang calls object routers. These are special-purpose network appliances that support reliable, high-performance Internet messaging.

"We have tremendous expertise, not just in software, but in making the Internet work better," says Tuttle. In addition to scalable message queuing, the object routers

optimize packet flows across Internet backbones to avoid congestion, and ensure that sessions will remain persistent despite unpredictable Internet connectivity.

Bang customers tag objects in their Web pages that they want to have updated in realtime, similar to the way they tag graphics to be retrieved through content-delivery networks such as Akamai. Bang updates only those portions of the page through its own infrastructure, eliminating the need to refresh the whole screen for every change.

When Bang launched in April 2001, its initial customers were online media companies such as CBS Sportsline and Excite@Home. These companies used Bang's network to update content such as sports scores and instantly deliver it to large audiences of users. As content providers faltered amid the dot-com shakeout, Bang shifted its focus. It found that a large percentage of its prospective customers (especially the ones likely to be able to pay!) were in the financial-services sector, so it decided to make that its primary market focus.

BANG NETWORKS INFO

Headquarters: San Francisco, CA
Founded: April 1999
Employees: 40
Funding: $32 million from Sequoia Capital, Atlas Venture, Marc Andreessen, Dan Warmenhoven, Nicholas Negroponte, Jeff Parker, Angel Investors LP and Wilson Sonsini Goodrich & Rosati
URL: www.bangnetworks.com

Bang's customers include Dow Jones, S&P, the Federal Reserve Bank and GovPX, which is the primary provider of pricing information for US Treasury securities. When GovPX lost its primary network connectivity after the September 11 attacks, it quickly switched over to Bang's infrastructure to deliver pricing information rapidly and reliably to its customers.

Speed and reliability are paramount in financial services. A nanosecond can make the different between a profitable trade and an unprofitable one. That's one reason it has been among the biggest markets for LAN-based messaging buses such as Tibco. Financial-services companies also value global reach. They may have trading floors in New York, London, Tokyo and other cities, all of which must have access to the same information at the same time. Bang's network-based approach works well in this environment.

Bang originally offered its technology purely through a service model, using its own network. Large financial-services firms often have their own global infrastructure, so Bang now also licenses its software to deploy into customer networks.

KENAMEA INFO

Headquarters: San Francisco, CA
Founded: July 1999
Employees: 75
Funding: $41 million from Crosspoint Venture Partners, Azure Capital Partners, Lightspeed Venture Partners and Palm Ventures
URL: www.kenamea.com

Kenamea: hotwiring the front end to the back end

Kenamea has developed a software platform for realtime applications that can be installed adjacent to an enterprise application server. As ceo John Blair explains, "It sets up a transactionally safe, fully secure and bi-directionally event-driven application network that hotwires together the application server and end-user devices or other application servers."

Blair believes one of Kenamea's important selling points is that it fits directly into the existing enterprise application development infrastructure: "You can take today's application servers and today's front-end environments and you can build true Net applications." Kenamea stitches together the HTML/XML interpreter that drives the user interface with the back-end logic in the application server (such as an Enterprise Java Bean or a Microsoft Distributed Common Object Model object).

On the front end, Kenamea gives customers two options. It can work within the browser, or for applications that require a complete security model and transactional integrity, it automatically downloads a client-side extension. The extension depends on the operating system; on Windows it's an ActiveX control.

A typical application for Kenamea is within a Regional Bell Operating Company (RBOC) that is required under the 1996 Telecommunication Act to provision wholesale local lines for competitive local exchange carriers (CLECs). With existing systems, it takes 10 to 42 phone calls and faxes to provision each line. Using Kenamea's technology, the CLECs can see directly into the RBOC's back-end systems, and can dynamically provision lines over the Internet. In this and other cases, Blair says, interactivity is the primary value customers get from Kenamea. Reliability, transactional integrity and security are basic requirements to get in the door.

In the current spending environment, most Kenamea customers are focused on improving operational efficiency and reducing costs, Blair says, rather than topline growth. Thus, most initial Kenamea deployments are for internal productivity-based applications. However, the company sees some customers and prospects looking to cross-organizational applications such as the telephone example above.

Kenamea recently received a grant of $200,000 from CommerceNet as part of its next-generation Internet effort. Kenamea is using the money to fund professional

services in connection with cutting-edge customer deployments, including an investment bank trading-room dashboard and a mobile order-management system.

Palm invested in Kenamea because it wanted to integrate its wireless-enabled devices with enterprise applications. Wireless environments are inherently unreliable. Because it's impossible to guarantee continuous connectivity, systems must offer caching and synchronization. Kenamea offers this capability built into its system.

KnowNow: application internetworking

KnowNow debuted at the 2001 PC Forum (SEE **RELEASE 1.0**, MARCH 2001). Since that time, it has filled in its management team, including ceo Bill Barhydt, who founded online collaboration vendor WebSentric after serving as an executive at Netscape. The company has also sharpened its focus on the enterprise market, labeling what it does as "application internetworking."

KnowNow works entirely within the standard Web infrastructure of HTML/Javascript on the browser, HTTP for transport and uniform resource names (a generalization of the Web's URLs) to identify resources. It uses small Javascript micro-servers that run on the client machine and trick the browser into keeping HTTP connections open persistently. On the server side, it has software it calls event routers that manage the flow of messages among clients.

KnowNow's approach allows it to deliver realtime many-to-many application performance without any software installation, because the 30K micro-servers download and run transparently. Its event routers allow for data transformation, content manipulation or business logic to act on data in the middle of the event stream. With a microserver on each client, it can treat every transaction the same, whether it originates from a Web page or an enterprise application, a browser or an Excel spreadsheet, inside or outside the firewall. "There's no difference between making information visible via the intranet, via the extranet, or via the Internet," says Barhydt.

KnowNow's big advantage, Barhydt argues, is that it radically reduces the cost to implement realtime visibility. Because it uses existing standards all the way through, none of the participants in an application network need to deploy anything new. Instead of requiring a commitment to enterprise-wide deployment and pressure to get everyone to implement the software, KnowNow can easily be installed in a department and expanded incrementally. Developers can download tools and trial

software from KnowNow's Website to experiment with realtime asynchronous communications, or to extend KnowNow's framework to new endpoint applications.

Realtime is on Our Side

The competitive neighborhood

Realtime computing ties into the broader trend toward Web services. Companies are shifting from proprietary, tightly coupled integration frameworks such as COM and CORBA to open standards such as SOAP. If you think of Web services as a movement to re-conceive software development around Internet standards, the three companies described above represent the transport layer underneath the XML-based components.

Other companies such as FineGround, Sockeye and Peribit address speed and reliability of sending data across the Internet through content acceleration and network traffic analysis. Venation, a recent spinoff from British Telecom with an investment from Softbank, is also in this space. Unlike content-delivery networks, which improve Internet performance by serving objects from local caches, these services understand and transparently manipulate the content itself to optimize services.

An adjacent set of startups, including Grand Central, Flamenco Networks, Slam Dunk Networks and TransactPlus, puts more emphasis on security and reliability, and on enabling server-to-server (as opposed to user-facing) transactions across Web-services networks (SEE **RELEASE 1.0**, OCTOBER 2001). Another group, including PowerMarket (SEE **RELEASE 1.0**, MARCH 2001), SeeCommerce and Tilion, delivers realtime XML-based analytics on top of application networks for supply chains and trading partners. Finally, vendors such as Crossweave, Cacheon, Asera and iSpheres help companies create composite applications that integrate data and functions from disparate enterprise systems in realtime and present them to users through common interfaces.

Consolidation and partnerships in this area are inevitable, both among these companies, the realtime vendors and the composite application integrators, and between startups and larger platform providers. Microsoft, for example, has outlined a global XML Web-services architecture (GXA), including proposed standards for dynamic routing of Web-services messages. However, though Microsoft has announced or shipped many of the pieces through its .Net effort, it doesn't yet have an Internet-scale event-based messaging platform.

The right time for realtime

The initial benefit of Internet-based realtime computing is cost reduction. Private networks are expensive, and so are the messaging buses that enterprises use today to run their internal mission-critical applications. In today's economic climate, there's a strong selling proposition for technologies that can deliver the responsiveness, reliability and security of these networks for a fraction of the cost over the Internet.

Eventually, realtime computing will have deeper consequences. Users will be able to see instantly what's happening across an enterprise or a network of business partners. That will make for more decentralized, more responsive businesses. "We're moving toward dynamically configurable businesses, where information itself is always on," says KnowNow's Barhydt. "When information changes, it tells you that it's changed, as opposed to waiting for you to ask whether it has changed."

Such visibility is incredibly powerful. Business intelligence and supply chain management (SCM) software were supposed to deliver something along these lines, but they have come up short in high-profile cases. Nike, for example, blamed SCM vendor i2 for an inventory-related earnings shortfall at the beginning of this year.

Even Cisco, legendary for its cutting-edge IT infrastructure, was stuck with billions of dollars of excess inventory when demand for its routers fell suddenly. In Cisco's case, the problem may have been the company's failure to respond to the information it had. Even so, making more processes and data visible more quickly to more employees can only help.

Realtime computing means that business participants can see and interact with decisions as they happen, rather than relying on prospective forecasting or retrospective analysis. Those tools have their place, and it's unrealistic to expect every process to be accessible in realtime, but providing direct connections to high-value applications can generate significant benefits.

For example, consider a Kenamea customer that manages money for institutions. Today, it sends those institutions an email at the end of every day describing the trades it executed pursuant to each institution's defined strategy. Using Kenamea, the bank can give the customers a realtime dashboard showing the transactions as they happen, and giving the customer the opportunity to become involved. "The customer is now much more aware, much more empowered, much more plugged into the realtime decisions being made on their behalf," says Kenamea ceo Blair.

Realtime computing will also affect the business aspects of enterprise applications. One reason so many application service providers (ASPs) failed is that they took traditional applications designed for local operation and simply slapped Web front-ends onto them. Users were faced with slow, unfamiliar interfaces and unreliable performance, and the ASPs themselves had to build out huge data centers and networks to support these jury-rigged applications on the back end. Realtime computing provides the reliable high-performance messaging platform that a new generation of ASP services could ride on top of, overcoming these limitations.

"The Internet today is where the LAN was in the 1980s," before middleware and messaging buses swept through the enterprise computing environment, observes Bang Networks' Tuttle. Over the next few years, Web services and realtime computing will have at least as significant an impact, only this time the sandbox is bigger.

Looking to the future, Blair waxes enthusiastic: "If the browser enabled the first generation of the Web, and the application server enabled the second, then Internet-scale realtime communications will enable the third – true network applications that are dynamic and fully interactive, fast and trusted. That will rock." ■ R 1.0

New Adventures in Text Processing

BY KEVIN WERBACH

There's more to text than words. Humans pick up on concepts and latent structure to understand what documents mean; for example, whether a person is the actor or the object of an action, or whether the use of Bill and Java in the same sentence means that Bill knows Java. . . or that he is an expert who wrote the book. Computers can do wonderful things when information is organized into fields that can be manipulated in databases, but they have a hard time with ordinary unstructured documents.

Intelligent text processing uncovers the patterns in documents. Using a variety of techniques, systems can extract structure, classify information or allow searching on fuzzy concepts. These capabilities, in turn, can serve as the foundation for other applications ranging from corporate portals to job applicant tracking. As XML and Web services take hold, programmatic access to textual material becomes even more important, because these mechanisms are based on structured documents.

Text processing is among the best-trod fields in computing. It's one we've covered frequently over the years (for example, **RELEASE 1.0,** JULY 1991 on SGML extraction and **RELEASE 1.0,** APRIL 1991 on text retrieval standards, plus earlier issues). Work in natural language processing and machine learning has yielded powerful algorithms to recognize concepts in documents. Many companies in fields such as document management and knowledge management have put these techniques to work to make it easier to organize and search for information (SEE **RELEASE 1.0**, JUNE 2001).

The area continues to attract interest for a simple reason: There are an awful lot of documents out there. "Maybe the major defining trend of our time is the proliferation of unstructured information, and the breakdown of tools for managing unstructured information," observes WhizBang vp of strategy Barney Pell (SEE PAGE 15). The Internet made information overload even worse, by putting everything on a common platform. As Mohomine ceo Neil Senturia (SEE PAGE 19) colorfully puts it, "We're squirrels. If you don't automate this process, you'll drown in nuts."

There are many ways to manipulate and manage unstructured information. In general, *parsing* means processing a document to identify its structural or conceptual elements; *retrieval* means bringing back the most appropriate documents for a particular request; *extraction* means pulling out fields of information which can be

plugged into a database or other manipulable format and then searched, managed or fed into another application; and *classification* means assigning documents to categories or topics in a taxonomy. But the lines between the functions are blurry, and sometimes one mechanism is just an intermediary step to another.

Below we discuss several startups that, using techniques derived from artificial intelligence (AI), have developed new approaches to information retrieval and extraction. Unlike earlier vendors, these companies don't require pre-scripted lexicons or rules for a specific type of document. They can identify concepts and fields in new documents by analyzing example files, recognizing the patterns automatically.

Sense vs. search

Many companies are working to better handle unstructured data. Web-wide search, for example, has improved significantly since Google came on the scene (SEE **RELEASE 1.0**, JANUARY 1999). Searching the Web, however, is a different task from sifting through documents on an enterprise network, which most of the companies below concentrate on. The Web is vast (Google claims 1.6 billion documents indexed) and constantly changing, though it has suggestive hyperlinks and HTML structure. Using the Web is a matter of finding documents that are relevant for your query, and then reading those files. In unstructured data management in the enterprise, our focus here, the goal is different: to *do* something automatically with the documents.

Imagine a resumé-processing example. If you're filing a position for a Java developer and search a pile of resumés for the word "Java," a keyword search engine will give you a list of documents. Some will be people who list "Java" under skills but have no actual experience; some may have worked in marketing in the Java group at Sun or at a coffeehouse called "Java Station." And the best candidate may have referred to "Visual J++" in her resumé instead of "Java," and not shown up in the results at all.

Moreover, sometimes finding matching resumés isn't the point. What you really want may be to identify candidates who have three years' experience as software developers, Java knowledge and who live within 20 miles of your company. Or you may want to feed in your job description and ask for a ranked list of the candidates who best match the listed requirements based on their resumés. You may want to classify resumés into a hierarchical taxonomy, pulled into a corporate portal interface (SEE **RELEASE 1.0**, FEBRUARY 1999). Or you may want to feed them into an applicant-tracking system that requires a series of tagged fields.

Those aren't search functions. They are questions that require some understanding of how the documents are structured and what they mean. Mohomine's Senturia defines the challenge as, "Can you make the application smarter by learning how to structure all the unstructured data that, until now, was available only to the application manually?"

Resumé masters

We used an example above from human resources (HR) for a reason: All of the startups described below have a strong beachhead in that market. Says Mohomine's Sameer Samat (SEE PAGE 19), "Resumés are a perfect example of a semi-structured or unstructured piece of textual data." The Internet has transformed the hiring process, allowing millions of people to post resumés on sites such as Monster.com. However, all those resumés (Monster has 12 million) quickly become overwhelming.

A conventional keyword search engine can match a formal criterion – types 50 words per minute, lives within 50 miles of Chicago – but that can generate an avalanche of responses. And ultimately what the employer wants isn't the word "marketing" on a resumé, but a sense of whether the candidate's background and experience suggests proficiency and, just as important, a good fit for the specific job.

Hiring is a particularly good market for next-generation search and classification technologies because accuracy is money. An executive recruiter may generate tens of thousands of dollars in commissions for a placement, so finding the right person for a job is worth paying for. Human resources functions aren't optional; the question is how efficient and productive organizations can make them. And because there's such a direct nexus to compensation (whether of an employee or a recruiter), companies in the hiring space are always looking for ways to differentiate themselves. Even in a down market, therefore, new vendors can find customers in this area.

Moreover, the hiring market has shifted dramatically with the economic downturn. Where in the past few years employers went begging for qualified candidates, today every job listing gets a deluge of applications. Sifting through those applications and finding the right one thus becomes especially important.

Resumix, started in 1988, was the first to apply AI techniques to classify, search and retrieve resumés, but it required pre-defined definitions for each type of document or job. The new companies discussed below can create their models automatically and update them on the fly as new documents are added. Burning Glass is the only

THE SPOOKS ARE LISTENING

Since September 11, law enforcement and intelligence agencies have stepped up their interest in information extraction. These organizations need to sift through vast troves of information to identify criminal or suspicious activity.

The CIA-sponsored venture fund In-Q-Tel is an investor in both Mohomine (SEE PAGE 19) and Stratify (SEE **RELEASE 1.0**, MARCH 2001), and through that relationship both companies have government deployments they can't talk about. Engenium (SEE PAGE 16) was recently selected as the search and classification engine for the US Defense Department's defense messaging system (DMS) product, which is being developed by California Microwave Systems, a Northrop Grumman subsidiary.

one focusing exclusively on HR today; its strategy is to go deep – predictive HR applications – rather than broad. Engenium and Mohomine get most of their revenues in HR (including OEM deals) but are pushing into other markets. WhizBang built its first implementation as the back end for the Flipdog.com job search site, but the bulk of its customers now are content providers and others.

The Sense-Makers

Like others who offer intelligent text-processing technology, the four companies we cover below take advantage of a basic feature of text: Words aren't distributed randomly. Documents have certain regularities which can be picked up through statistical or other computational techniques. A computer doesn't really know that a business letter is different from a resumé, but it can be trained to spot the telltale structural differences. A letter, for example, usually starts with a date, the recipient's name and address, and a salutation.

It's fairly easy to hard-code programs that look for specific features – Microsoft Word, for example, automatically recognizes documents that look like business letters and asks if you want to apply automatic formatting templates. It gets much harder if you want the computer to figure out the structure of a type of document it's never seen before. The problem becomes one of machine learning, where the user *shows* the system a group of examples rather than *telling* it what to look for.

Naturally, every company claims that its algorithms are the best, that it has hired the top academic experts and that it's the first to develop an extraction engine suitable for any type of document. Ultimately, the proof is in the customers.

Each of the four startups below takes a slightly different approach. WhizBang combines extraction and classification with spiders that crawl the Web and pull back chunks of information to create or enrich local databases. Engenium, the only one

that doesn't do extraction, uses machine learning to understand documents so they can be retrieved or classified based on concepts instead of keywords. Burning Glass builds analytical and predictive models about candidates' likely job performance. Mohomine offers information extraction as a component for other applications.

WhizBang Labs: super duper pattern recognition

WhizBang founder and cto Dallan Quass was a co-founder of Junglee (SEE **RELEASE 1.0**, MARCH 1998). Junglee was one of the first companies to develop scraping software that could pull information such as prices and product listings from Websites and incorporate it into aggregated databases that could drive comparison-shopping applications. However, it required hard-coded programs for each site.

With WhizBang, Quass set out to build technology that could automatically adapt to any page and understand its information structure. "For us, information extraction is about creating that structured base of content," says vp of strategy Barney Pell. The technical challenge was to build machine-learning algorithms that could accurately identify patterns in different kinds of Web pages.

As a proof of concept, WhizBang's team created Flipdog.com, a job-listings site that automatically pulls in job announcements from 50,000 corporate Websites and creates a searchable database that job searchers can drill down on. This gave the company valuable real-world experience with its technology. In May, WhizBang sold Flipdog to TMP Worldwide, which operates the Monster.com job board, with a subsequent agreement to provide back-end technology to all of TMP's sites (soon to include HotJobs, pending regulatory approval). WhizBang now provides its software and services as infrastructure to other companies.

WhizBang's next set of customers were content vendors, who recognize the benefit of finding and making visible the underlying structure of documents because it's a central part of their business. Dunn and Bradstreet uses WhizBang to augment its business information databases. WhizBang pulls data such as addresses, phone numbers and corporate officers off Web pages, and puts them into the proper field in the D&B database. D&B also uses the system to perform analysis such as finding all the shopping carts on the Web and which credit cards they accept. It can then sell that information to customers interested in selling services to online retailers.

WHIZBANG INFO

Headquarters: Provo, UT
Founded: March 1999
Employees: 140
Funding: $35 million from Dominion Ventures, UBS Capital, HP, Bear Stearns, Spiral and Stars, InterVest, New Media Venture Partners and individuals
URL: www.whizbang.com

Next, the company hopes to gain traction in the enterprise market. Pell sees demand for several kinds of uses, including identifying sales leads, text transformation and adding XML to unstructured documents, building on the trend in enterprises towards driving all applications from databases.

Pell says one common form of competition for WhizBang is manual coding of documents, usually by teams of workers in low-cost markets such as India. Though WhizBang believes its automated approach is faster and reliable, it is developing its own center in India to use its tools to augment manual coding for those customers that want such an approach.

Engenium: the genius of conceptual search

Engenium was started three years ago by chairman Dave Chaplin, who was building a knowledge management system at KPMG, and president Dave Copps, who approached Chaplin as a vendor of realtime information services. The two shared concerns about the state of the art in enterprise search technology, and decided to do something about it. "Most of these systems bottlenecked at a keyword engine," says Copps. "We were looking for a way to build a technology that would dynamically learn as information entered the system."

ENGENIUM INFO
Headquarters: Dallas, TX
Founded: 1998
Employees: 15
Funding: internal
URL: www.engenium.com

The result was conceptual search technology called Semetric that identifies concepts from patterns of words and uses those to build an index called a ConceptSpace. Unlike a simple word index, the ConceptSpace associates different terms that refer to identical or similar concepts. Each time a new document is put into the system, its words are mapped against the ConceptSpace.

Unlike WhizBang, Burning Glass and Mohomine, Engenium doesn't try to extract structured fields from documents. Engenium believes that conceptual search should be the starting point for analyzing unstructured documents, because if you can search on fuzzy concepts directly, you don't need the structure. "We see extraction as something that should follow the matching process," says Copps. Even the best extraction technology is only 90-percent accurate, he notes. Of course, whether that's good enough depends on the application.

The main advantage of Semetric is that it doesn't rely on keywords. "Conceptual search is treating small groups of terms or even individual terms as amorphous. Their meaning is defined as a function of the context in which they are held,"

CLASSIFICATION AND TAXONOMIES

As we discussed in our knowledge management report (SEE **RELEASE 1.0**, JUNE 2001), taxonomies are an effective way to organize information into hierarchical categories, but manually creating them is cumbersome.

Stratify, formerly Purple Yogi (SEE **RELEASE 1.0**, MARCH 2001), is one company that automates the organization of internal and external documents into hierarchical taxonomies. Over the past nine months, the company has positioned itself away from classifying news-feeds and towards helping enterprises to manage information that doesn't fit neatly into the black-and-white world of databases. Different parsing techniques work better or worse depending on the type of document, ceo Nimish Mehta points out; what's ultimately important is what you can do with the text the system has manipulated.

"Why do you want to classify text?" Mehta asks rhetorically. He continues: "It used to be, so you could find things. That's good and a useful problem. But what if I want to use this and build applications around it?" Stratify sees classification as a core technology for the emerging industry of unstructured data management.

An interesting classification competitor is Quiver, which began life as a directory-building tool for consumer-facing vertical portal sites. As that market dried up, Quiver refocused on enterprise taxonomies. What makes the company unique is the way it combines automated mechanisms with human beings through a workflow process. "We fundamentally believe [taxonomy-building] can't be solved algorithmically," says ceo Scott Potter.

Quiver does automatic classification, using technology its founders developed in Israel, but only as a first cut. It provides workflow tools to route content to editors within a company who make the final determination. The system uses on fine-grained business rules, so that, for example, a customer could have all classifications automatic in an external news-feed topic, those falling below a confidence threshold in a competitive intelligence topic routed to editors and everything in a sexual harassment section of the HR intranet reviewed by a human.

Says vp of sales and marketing Andy Feit, "Most, if not all, of our competition are black boxes. They do what they do, they put content into buckets at anywhere from 60-to-80 percent accuracy." Many customers want something as good as a human-edited taxonomy, but they want to reduce the time and cost of getting there. That means they need flexibility to override the automated systems. "Once Autonomy puts something in a topic, there's no easy way to get it out," says Feit.

In the current market, Quiver sees companies pulling back on massive corporate portal projects, but still going forward with taxonomies as an incremental way to organize their internal and external content.

explains vp of r&d Matt Sommer. Engenium combines probabilistic mechanisms (the likelihood of a term occurring with others based on prior observations) and forecasting models that represent concepts as vectors in multi-dimensional space.

The first application based on Engenium's technology is HireReasoning, for resumé processing. The company signed up many of the biggest customers in executive recruiting, including Korn Ferry's FutureStep, Heidrick and Struggles, and staffing firm Spherion. About three months ago, Engenium began to push into other markets, with the Defense Department OEM deal (SEE PAGE 14) being its first major success. Engenium hopes to partner with leading providers in other markets, including enterprise portals, which could embed its technology.

Speed of deployment is a big attraction for Engenium's customers. With Korn Ferry, for example, Engenium was able to make FutureStep's database of 250,000 executive profiles searchable within less than four hours. For some customers that want fast deployment, Engenium ships a "starter brain" ConceptSpace that teaches the system

patterns in a class of documents, such as resumés, even though the customer won't actually be searching on the example resumés.

Engenium has about 20 customers today. Some of these are server licenses, starting at about $150,000 for a two-processor deployment, while others are OEM deals which generally include an up-front fee of $50,000 to $500,000 and a percentage of back-end sales revenues.

Burning Glass: resumé scoring

The founders of Burning Glass came from HNC Software. There's a connection: HNC has a significant business finding patterns in vast pools of transactions and using them to generate individual predictions about credit card fraud, creditworthiness, customer profitability and so forth. Burning Glass has created algorithms to find patterns in vast pools of resumés. "We do not view a resumé as a bag of words. Instead, we view it as a history of a career in progress," says ceo Krishna Gopinathan.

The company first built a database of 20,000 resumés with fields manually assigned by offshore workers. It used natural language processing to train its extraction engine to recognize over 400 elements within a document. In addition to the raw fields such as job titles, it interprets much of the data – for example, adding SIC codes to categorize the kinds of companies someone has worked for. After extracting information from the training documents into its KnowledgeMine, Burning Glass applied predictive modeling techniques to a larger set of over 500,000 resumés, identifying patterns in how employees move from job to job and perform within those jobs.

BURNING GLASS INFO
Headquarters: San Diego, CA
Founded: October 1999
Employees: 10
Funding: $4.7 million from HNC Software, former Tandy ceo Luther Henderson, Pier One founder Arjun Waney, Qualcomm ceo Irwin Jacobs
URL: www.burningglass.com

Burning Glass sells information-extraction solutions, and that's what its initial customers have bought. But more exciting are the offerings it builds on top of extraction, which Gopinathan calls "predictive employee relationship management." The first such engine identifies how well a candidate matches a job. Burning Glass doesn't just score matching terms, but how likely it is a candidate will be placed in a job based on previous examples. Customers can use the scores to narrow down or rank candidates.

The latest module is longevity scoring. "We look at the resumé data, and we're able to say, 'given this job posting and this resumé, this person is likely to stay in this job x number of years'," explains Gopinathan. This sounds like science fiction, but

CAESIUS: THE WEB AS A DATABASE

WebQL, a software product from Seattle-based startup Caesius, allows users to pull structured data from the Web directly using a derivative of SQL, the standard query interface for enterprise databases. In a sense, it's a hybrid of information extraction (which automatically pulls structure out of documents) and search (which actively retrieves documents based on user queries).

Caesius calls this process Web harvesting or Web farming. Perhaps the most common uses for it are competitive intelligence and market research. Many companies use home-grown scripts to crawl the Web and retrieve specific types of information. Caesius packages this capability into a general-purpose software package. The company delivers its software as a shrink-wrapped package, as turnkey software customized for a particular use or as a hosted service.

Gopinathan says the company has been able to back up its claims by back-testing them on historical data. For the five percent of resumés scoring highest in longevity, 92 percent actually stayed more than three years, while for the five percent scoring lowest, 86 percent stayed less than three years.

"We score employees and candidates analogous to the way people get scored for credit," Gopinathan says. When you apply for a loan, a bank gives you a credit score based on your past activities. Similarly, Burning Glass looks at factors such as how many previous jobs a candidate has had over a given period of time, his or her longest and shortest previous job and the skill match between previous jobs and the current one. Surprisingly, too close a match in skills is a negative indicator for longevity. People get bored if they do the same thing for too long.

Burning Glass customers include Tapestry.net, Guru.com and Volt, which use the extraction product. It offers its technology either as licensed software that customers can run internally, or as a hosted service where it charges for each resumé parsed. The extraction service costs $1 to $2 per resumé, based on volume. Longevity scoring will be about $20 per resumé. Upcoming predictive modules include macro-level supply and demand analysis, promotability and workforce management analytics.

Gopinathan says that, after a round of layoffs bringing it down from over 20 to 10 employees, the company is profitable. Its existing investors have expressed willingness to provide additional funding if needed, he adds.

COMING SOON

- Online collaboration tools.
- P2P grows up.
- Distributed distribution.

- And much more. . . (If you know of any good examples of the categories listed above, please let us know.)

Mohomine: extraction infrastructure

Mohomine's four founders are computer science graduates of the University of California, San Diego. Beginning in 1998, they developed a Website, Sourcebank.com, to identify and organize online software-development resources. When they launched the site in

August 1999, they began receiving inquiries about whether they could use the same mechanisms to classify data in other industries. "We had set out to solve a very specific problem, but what we began to realize was there was a much larger opportunity," says co-founder and cto Sameer Samat. The founders linked up with ceo Neil Senturia, who had recently sold high-speed access provider Atcom to CAIS, and in January 2000 they launched Mohomine.

MOHOMINE INFO

Headquarters: San Diego, CA
Founded: January 2000
Employees: 33
Funding: $10 million from In-Q-Tel, Windward and Hamilton
URL: www.mohomine.com

Mohomine's solution isn't intended directly for end-users; it is an infrastructure layer that plugs structured XML data into applications. "People lay out information in a variety of different ways, because they are trying to convey information to other people. Machines are used to seeing information lined up in rows and columns of databases," says Samat.

For example, PeopleSoft uses Mohomine to enhance the e-recruiting module of its flagship human resources suite, to manage resumé processing for job applicants. Mohomine extracts over 250 attributes from resumés automatically, including personal data, employment history, education, skills, certificates/membership and miscellaneous information such as career objectives.

Though Mohomine's initial deployments are in HR, the company says its technology works equally well in other contexts. The learn-by-example mechanism isn't specific to just one type of document; to apply the algorithms to a different vertical just requires training the system with a new set of examples. Mohomine says it is in pilots or close to signing deals with customers who will use its technology in customer relationship management, document management and enterprise portals.

The New Information Toolbox

Overcoming the information deluge is a multi-faceted challenge. Extraction, classification and conceptual search are useful tools, but hardly the only ones. In addition to Verity and Autonomy, the established vendors in enterprise search and classification, newer players are coming on the scene such as Albert, which learns from each user's previous queries to better understand what that user is looking for, and ClearForest, which summarizes documents and visually maps their relationship to one another.

Resources & Contact Information

Tim Tuttle, **Bang Networks**, 1 (415) 551-1201; fax, 1 (415) 934-2951; ttuttle@bangnetworks.com

Krishna Gopinathan, **Burning Glass**, 1 (619) 723-1891; krishna@burning-glass.com

Kelvin Chin, **Caesius**, 1 (206) 443-6836; fax, 1 (206) 269-0694; kelvin@caesius.com

Dave Copps, Dave Chaplin, **Engenium**, 1 (469) 374-9464; fax, 1 (469) 374-9469; dave.copps@engenium.com, dave.chaplin@engenium.com

John Blair, **Kenamea,** 1 (415) 869-5012; fax, 1 (415) 869-5005; john@kenamea.com

Bill Barhydt, Rohit Khare, **KnowNow**, 1 (650) 641-8500; fax, 1 (650) 641-8502; bill@knownow.com, rohit@knownow.com

Neil Senturia, Sameer Samat, **Mohomine**, 1 (858) 362-3000; fax, 1 (858) 362-3099; neil@mohomine.com, sameer@mohomine.com

Scott Potter, Andrew Feit, **Quiver**, 1 (415) 863-9945; fax, 1 (415) 863-9946; spotter@quiver.com, afeit@quiver.com

Nimish Mehta, **Stratify**, 1 (650) 210-2713; fax, 1 (650) 938-5716; nimish@stratify.com

Barney Pell, **WhizBang**, 1 (801) 418-7100; fax, 1 (801) 818-0300; bpell@whizbang.com

In the past, text extraction and classification were primarily done so that users could search on fields rather than a word index. The question is whether these functions will remain an ingredient for other applications (as enterprise search engines are today), will establish themselves as a sort of content middleware (taking unstructured information and transforming it so that it can be processed in databases), or, most intriguingly, will be the foundation for a separate software category: unstructured information management.

The argument for a new category is that companies do different things with unstructured data than with structured data. Organizing this information is the first step, as it was with databases. Once that piece is in widespread use, vendors will develop higher-order applications on top. Burning Glass' predictive resumé scoring is one example of the kinds of application that information extraction enables. Because of the accuracy and flexibility they offer, and the prevalence of common Internet standards for documents and applications, the new vendors can tackle a broader market.

"Because the tools haven't existed, the IT departments haven't thought about the applications they could create," says Stratify's Mehta (SEE BOX, PAGE 17). That will change. In the structured data world, relational databases became the foundation for everything from business intelligence to financials to sales force automation. Watch the unstructured space to see whether a similar ecosystem emerges. ■ R 1.0

Calendar of High-Tech Events

2002

JANUARY 7-11	**MACWORLD CONFERENCE & EXPO** - San Francisco, CA. The most comprehensive gathering for the Macintosh community. Register online, or call 1 (800) 645-EXPO for information. www.macworldexpo.com
JANUARY 8-11	**CONSUMER ELECTRONICS SHOW** - Las Vegas, NV. The place to be to see the latest, hottest consumer technology. Scheduled keynoters include Bill Gates, Bob Pittman and Carley Fiorena. Register online,email CESinfo@ce.org for help or call 1 (703) 907-7605. www.cesweb.org
JANUARY 15-17	**NEXT-GENERATION WEB SERVICES** - San Francisco, CA. Infoworld's three-day event on the future of Web services. Speakers include Federal Trade Commission's Sheila Anthony and BEA's Bill Coleman. www.eventreg.com/nextgen/nextgen_1.htm
JANUARY 28-31	**O'REILLY BIOINFORMATICS TECHNOLOGY CONFERENCE** - Tucson, AZ. Covers the application of computing power to biomedical research. For more information, visit conferences.oreilly.com/biocon/cfp.html. conferences.oreilly.com/biocon/cfp.html
JANUARY 29-FEBRUARY 1	**LINUXWORLD CONFERENCE & EXPO** - New York, NY. The leading Linux event, offering the best environment to explore and experience Linux/Open Source solutions, products and services first hand. Register online or call (800) 657-1474 for information. www.linuxworldexpo.com
FEBRUARY 4-8	**MILIA** - Cannes, France. Come play with the next generation of interactive entertainment; the expo features the latest interactive content and TV, games,leisure software, broadband and wireless entertainment. Register online or contact, Laurine Garaude,telephone +33(1)41 90 46 30; by fax 33(1) 41 90 44 70, or by email laurine.garaude@reedmidem.com www.milia.com
FEBRUARY 10-12	**DEMO 2002** - Phoenix, AZ. Check out this year's crop of new startups and gadgets! For information, call Anna Aquino at 1 (800) 633-4312, or email registrar@idgexecforums.com. www.idgexecforums.com/demo
FEBRUARY 20-23	**TED 12** - Monterey, CA. Richard Saul Wurman's final year running this highly regarded cross-cutting event. He promises, "Simply the greatest design conference there ever was." www.ted.com
MARCH 24-27	**PC FORUM** - Scottsdale, AZ. EDventure Holdings' premier conference, now in its 25th year. Look for a speaker list and more details soon on our Website at www.edventure.com/pcforum. Contact Daphne Kis with any questions at 1 (212) 924-8800 or daphne@edventure.com. E K

E Events Esther plans to attend.
K Events Kevin plans to attend.

Lack of a symbol is no indication of lack of merit. The full, current calendar is available on our Website, www.edventure.com. Please contact Irene Lawrence (irene@edventure.com) to let us know about other events we should include.

pcforum 2002: Back to the Frontier

March 24 to 27, 2002
The Fairmont Scottsdale Princess
Scottsdale, Arizona

For our silver anniversary, we will guide the industry's return to its pioneering spirit. We hope you can join us – and your peers – as we explore the future both realistically and enthusiastically.

Whatever the economic climate, technologies succeed or fail based on how they solve problems and enrich people's lives. Entrepreneurs and businesses did not build the economy by fearing risks, but by taking risks, creating innovations, working hard and being smart enough to beat the odds. Today, as we all re-evaluate the place of information technology in the global economy, the high-tech industry needs the courage to lead its customers forward.

Join us at the 25th PC Forum to reinvigorate yourself for the challenges and opportunities ahead.

Some of the topics we are working on for this year's event include:

- How to Feel Secure in an Insecure World
- Web Services as the New Platform
- Overcoming the Broadband Bottleneck
- Lessons Learned: Strategies for Success in the Post-New Economy
- How to Be Yourself in a World Wide Web
- Turning Media and Entertainment Upside Down
- Control Over Information: The Battle for the Soul of the Net

Our growing speaker list includes:

Stewart Baker, *Partner*, Steptoe & Johnson
Steve Ballmer, *President & CEO*, Microsoft*
Craig Barrett, *President & CEO*, Intel
Tim Berners-Lee, *Director*, WWW Consortium*
Maria Cirino, *CEO & Co-founder*, Guardent
Cathy David, *VP & General Manager*, Target.com
David Dorman, *President*, AT&T*
Rick Ellenberger, *President & CEO,* Broadwing
Monte Ford, *CIO & SVP, IT*, American Airlines
J.C. Herz, *CEO*, Joystick Nation
Frank Ingari, *Chairman*, Wheelhouse Group
Floyd Kvamme, *General Partner*, KPCB*
Joe Nacchio, *Chairman & CEO*, Qwest
Jerry Parrick, *CEO & Founder*, Yipes
Steve Perlman, *President & CEO*, Rearden Steel
Michael Ramsay, *Chairman & CEO*, TiVO*
Hillary Rosen, *President & CEO*, RIAA
Stratton Sclavos, *President & CEO*, Verisign
Eric Schmidt, *Chairman & CEO*, Google
Mike Volpi, *SVP, Internet Switching & Services Group*, Cisco
Jerry Yang, *Co-founder & Chief Yahoo*, Yahoo!
Ed Zander, *President & CEO*, Sun Microsystems*

*awating confirmation

We recognize these are challenging times. PC Forum is where industry leaders go to take stock and to engage the future, and this year we are offering a steeply discounted early registration fee of $3500. To take advantage of this offer, please register immediately by using the form on the back or by visiting our Website at **www.edventure.com/pcforum**. Remember, after January 11th, the registration fee will go from $3500 to $4500. Sign up now!!

pcforum 2002: Back to the Frontier

March 24 to 27, 2002
The Fairmont Scottsdale Princess
Scottsdale, Arizona

REGISTRATION (for more details please visit www.edventure.com/pcforum)

Your Information

Name _____ **Nickname** (for badge) _____

Title _____ **Direct phone** _____

Company _____ **Fax number** _____

Street address _____ **Your email address** _____

City, State, Postal code _____ **Company Web site** _____

Country _____

Assistant's name* _____ **Assistant's email address** _____

*We will send all email correspondence to both you and to the contact person or assistant listed above.

Attendee Web site

When we confirm your registration, we will issue you a secure password to the Forum Web site where you can communicate with other confirmed attendees. Please indicate your preferences below.

- ☐ Please post my name, title, company, email and company Web site on the private attendee site.
- ☐ Post only my name, company and company Web site on the attendee site.
- ☐ I prefer not to have any information posted on the site.

Registration Fee

- ☐ Full registration fee $4500
- ☐ **Early registration fee** (if received by January 11, 2002)..... **$3500**

Accommodations

Please provide your arrival and departure dates along with your room type preference below. EDventure Holdings will forward this registration form to the Fairmont Scottsdale Princess beginning in mid-January and until all available rooms are taken. The hotel will make the indicated reservation, charge your credit card for a one-night's room deposit and send you confirmation directly. Rooms from EDventure Holdings' room block are available to registered attendees only.

Arrival date _____ **Departure date** _____

Room type ☐None ☐Single $315/night (one king bed) ☐Double $315/night (two queen beds) ☐One-bedroom suite $540/night

Payment (Payment must be received with your registration form. However you will not be charged until we can confirm your attendance.)

☐ **Charge to my** ☐ Visa ☐ MasterCard ☐ American Express

Card number _____ Expiration date _____

Signature _____

Name and billing address on card (if different from above) _____

☐ **Check enclosed**** (Please make payable to EDventure Forums Ltd.)

**If you are paying by check and are requesting a hotel reservation, your credit card information is required to guarantee your first night stay.

Notice of cancellation must be submitted in writing to EDventure Holdings and is subject to a $500 service charge if received before February 15, 2002. Registrants who cancel after February 15 or who fail to attend are liable for the entire registration fee. Registrations to PC Forum are not transferable.

By signing below, you agree to abide by EDventure Holdings' cancellation policies and you further authorize EDventure Holdings to forward your room requirements and credit card information to the Fairmont Scottsdale Princess.

Your signature _____ Date _____